Automatic extraction methodology for accurate measurement of effective channel length on 65nm MOSFET technology and below

Dominique Fleury, Antoine Cros, Krunoslav Romanjek, David Roy, Franck Perrier, Benjamin Dumont, Hugues Brut

Automatic extraction methodology for accurate measurement of effective channel length on 65nm MOSFET technology and below

Dominique Fleury*†, Antoine Cros*†, Krunoslav Romanjek‡, David Roy*, Franck Perrier‡,
Benjamin Dumont‡, Hugues Brut*
*STMicroelectronics, 850 rue Jean Monnet, F-38926 Crolles, France
†IMEP, Minatec, 3 Parvis Louis Neel, 38016 Grenoble, France
‡NXP Semiconductors, 860 rue Jean Monnet, F-38926 Crolles, France
Email: dominique.fleury@st.com, Telephone: +33 (0) 438 923 314

***Abstract*—Constant downscaling of transistors leads to increase the relative difference between L_{mask} and L_{eff}. Effective length (L_{eff}) extractions are now crucial to avoid calculations errors on parameters such as the mobility, which can exceed 100% for shorter devices. We propose an industrially-adapted method to extract L_{eff} by using an enhanced "split C-V" method. Accurate and consistent values have been extracted ($\pm$1nm) and then correlated to mobility and HCI lifetime studies, as a function of L_{eff}.**

***Index Terms*—Effective channel length, split C-V, parasitic capacitances, gate-to-channel capacitance measurements**

I. Introduction

Transistor downscaling has been so fast in recent years that effective channel length (L_{eff}) – defined by the inversion layer length – hardly reaches 50% of the mask length (L_{mask}) on sub-65nm technologies. At such scales, a few nanometers shift can induce a misleading results interpretation, justifying necessity to estimate the channel length reduction ($\Delta L = L_{mask} - L_{eff}$) with enough accuracy, as a function of L_{mask}. The latter dependency results from deep sub-micron lithography limits ($L_{mask} - L_{poly}$ shift) and diffusion of Source-Drain Extensions (SDE) (Fig. 1).

As previously exposed, current-based extraction methods fail because of the mobility variations with the gate length [1], [2]. We developed an industrially-adapted method providing large scale extraction and so, statistical results thanks to fully-automatic probers. Technique has been improved to reach an unequaled 1nm-accuracy on L_{eff} (in relative) through a better understanding of parasitic capacitances. L_{eff} is a critical parameter, useful for Physics modeling and for a better understanding of the MOSFET. We will highlight this usefulness by examples of applications on mobility and HCI lifetime extrapolation as a function of L_{eff}, for 65nm-technologies and below.

II. State-of-the-art

$I_D(V_G)$-based methods have been previously proposed for L_{eff} extraction, assuming invariance of the mobility with the length of the transistor [3]–[6]. This approach is now mistaken because of the low-field mobility degradation observed on short devices [2], [7]. Capacitive method provides a L_{eff} extraction without any assumption regarding the mobility, but strongly depends on parasitic capacitances issues rising on modern technologies [2], [8]. As a consequence, we recently adapted the latter to sub-65nm technology devices, in an industrial context.

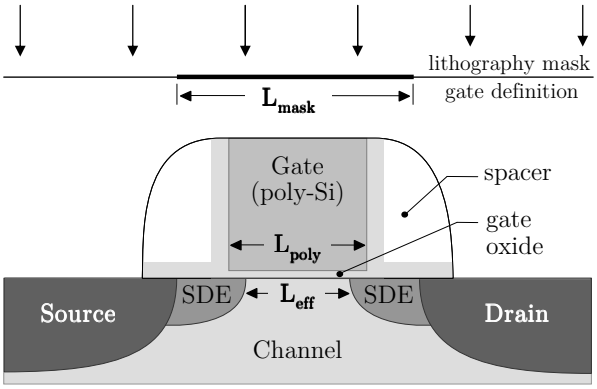

Fig. 1. Typical MOSFETs architecture studied in this paper – definition of L_{mask} (before SDE implantation), L_{poly} and L_{eff}

III. Methodology

A. Experimental setup

We performed 1MHz-frequency capacitance measurements on a fully-automatic 300mm-wafer prober (Accretech UF3000) equipped with a HP4284 LCR meter and an Agilent 4073B connection matrix. The latter is required for automatic measurements, allowing probing of several pad combinations. Specific home-developed software was used to perform batch extractions on large samples (20 dies per wafer) to improve accuracy and provide statistical results.

B. Capacitive method

L_{eff} is extracted using gate-to-channel capacitance measurement $C_{gc}(V_G)$ (Fig. 2) which is proportional to the channel area ($C_{gc} \propto W \times L_{eff}$). Furthermore, we will see that this method does not need any de-embedding structure to get

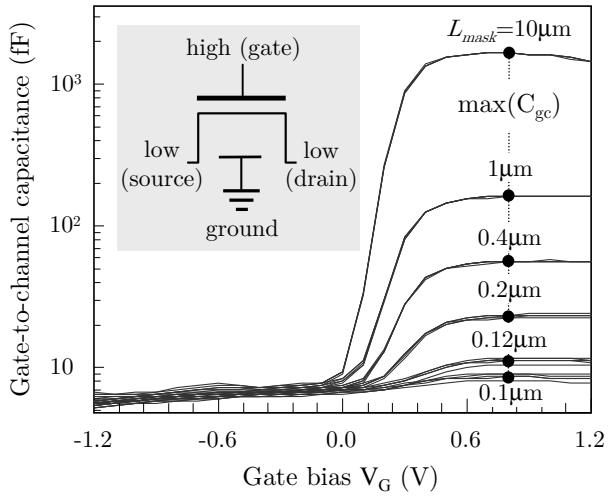

Fig. 2. $C_{gc}(V_G)$ curves measured for several transistor lengths and plotted using logarithm scale (45nm-technology, $T_{ox} \cong 12$Å). In insert: measurement setup.

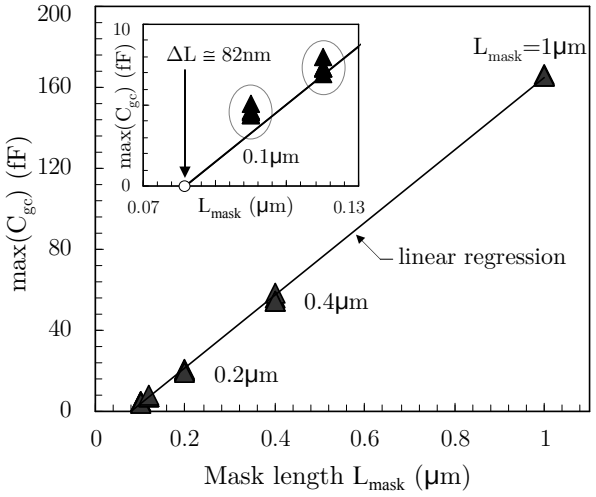

Fig. 3. Extraction of a constant ΔL from a linear regression on the plot of $\max(C_{gc}) = f(L_{mask})$. In insert: zoom on the specific part of the graph where ΔL is extracted

rid of parasitic capacitance as for $C_{gb}(V_G)$ measurements [1], [9]. Maximum of capacitance $\max(C_{gc})$ is set as a reference point for each curve. Two ways of extraction are thus possible:

1) Constant ΔL method: We can assume ΔL is invariant with L_{mask} (1) and extract its value from the linear regression on the plot of $\max(C_{gc}) = f(L_{mask})$ (Fig. 3) [2]. Thus, ΔL is the value read at the intercept between the linear regression and the L_{mask}-axis. In this case, error on L_{eff} will be strongly linked to the relevance of the $\Delta L(L_{mask})$ linearity assumption in the regression window. The latter is mainly influenced by lithography and gate-etch process optimization.

$$C_{gc} = W \cdot C_{ox} \times (L_{mask} - \Delta L) \tag{1}$$

2) Individual ΔL method: We can extract an individual ΔL for each transistor from a proportionality rule (2), using the longest transistor as reference (Fig. 4). Thus, the latter must satisfy to the relation $L_{mask}^{ref} \geq 1\mu$m in order to assume $L_{eff}^{ref} \cong L_{mask}^{ref}$ with enough accuracy. Error due to this assumption is not greater than 2% for a 1μm-length reference transistor (0.2% for a 10μm-length).

$$\Delta L^* = L_{mask}^* - \underbrace{L_{eff}^{ref} \times \frac{\max(C_{gc}^*)}{\max(C_{gc}^{ref})}}_{=L_{eff}^*} \tag{2}$$

Use of "individual ΔL" method allows extracting ΔL for each mask length and studying the $\Delta L(L_{mask})$ behavior due to photo-lithography and gate etch processes. Unlikely, the other method does not require a long transistor as reference but can only provide a constant ΔL which is almost the average value $\langle \Delta L^* \rangle$ (dotted line on Fig. 4).

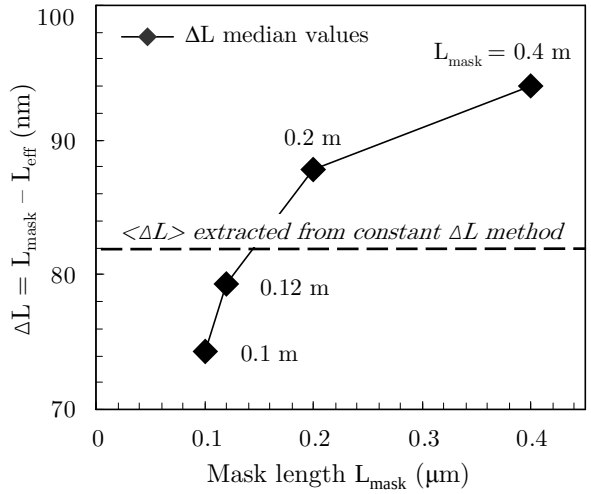

Fig. 4. $\Delta L(L_{mask})$ behavior resulting from "Individual ΔL" extraction method

C. Parasitic capacitance

Gate-to-channel measurement is impacted by parasitic capacitances: 1. a constant term comes from cabling, probes and connection pads; 2. a V_G-dependent component is inherent to the MOSFET architecture (Fig. 5). Total MOSFETs parasitic capacitance is composed by:

- the outer fringing capacitance (C_{of}) between the gate and source/drain through the spacers (this component does not depend on V_G);
- the inner fringing capacitance (C_{if}) between the gate and SDE through the channel;
- the overlap capacitance (C_{ov}) between the gate and SDE through the gate oxide.

In accumulation and inversion regime, C_{if} is screened by holes and electrons filling the channel. Its maximum value is thus expected near the flat-band voltage $V_G \sim V_{fb}$, when there is no screening possible. In strong accumulation, depletion region in SDE is created and acts as if the oxide thickness would have been increased in these regions, reducing C_{ov}.

Parasitic capacitance can not be measured directly in inversion regime. The latter has to be extrapolated to allow estimation of the parasitic capacitance behavior in inversion

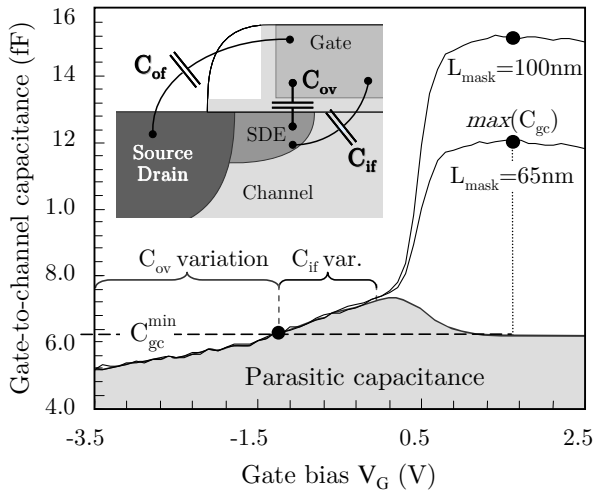

Fig. 5. Measured gate-to-channel capacitance and extrapolated parasitic capacitance for 65nm technology NMOS transistors ($T_{ox} \cong 18.5$Å, $W = 10\mu$m). In insert: schematic of intrinsic parasitic component of the MOSFET (from [10]).

regime (Fig. 5). Thus, we measure a C_{gc}^{min} parameter which has the same value (extrapolation) as the parasitic capacitance included in the $\max(C_{gc})$ measurement. Practically, $C_{gc}^{min} = C_{gc}\,(V_{th} - \Delta V)$ where V_{th} is the threshold voltage and ΔV is a constant adjusted from results in [10]. Typically, ΔL error generated by this procedure is not higher than 3% on nominal length transistors.

IV. Test structures

Capacitance measurements through the connection matrix require gate areas above 50μm^2 to provide a large enough Signal-to-Noise Ratio. Matrix test structures composed by N identical transistors wired together allow increasing the total area for a given L_{mask}. Thus, we used matrix test structures (Fig. 6) with constant width W and variable N to keep a near constant area ($A \cong 100\mu m^2$) whatever the length, providing to us the ability to perform automatic measurements in optimal conditions (2-3 min per $C_{gc}(V_G)$ curve). This test setup is fully-adapted and scalable to any Low Standby Power (LSTP) technology (gate oxide thickness $T_{ox} \geq 15$Å).

If $T_{ox} < 15$Å, the total area needs to be reduced further to get rid of the gate leakage effect which affects $\max(C_{gc})$ extraction [11]. Use of a connection matrix is no more possible but automatic measurements are still feasible by connecting the probes directly to the LCR meter. The measurement precision – and so the measurement time – have to be raised to keep the same accuracy. In all case, capacitance measurements on isolated MOSFETs are still possible if $W \sim 10\mu$m, considering the equipment detection limit ($A_{min} \sim 0.2\mu$m^2) which is reached for the shortest transistors.
For instance, a 45nm-LSTP nominal length transistor has an area about 0.4μm^2 ($W = 10\mu m$) and provides a signal amplitude of a few fF. For this critical case, the entire $C_{gc}(V_G)$ curve needs about 20min to be measured with enough accuracy.

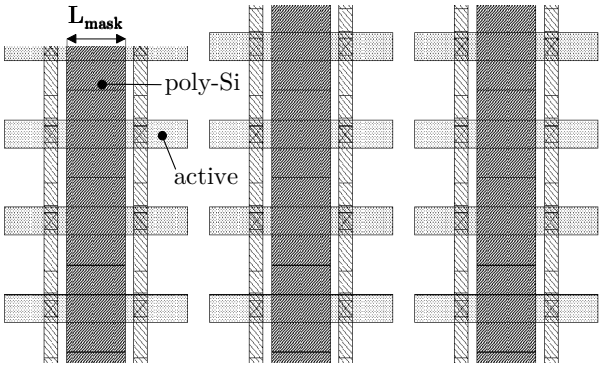

Fig. 6. Layout of matrix structures used for automatic measurements in optimal condition. $L_{mask} = 0.38\mu$m, $W = 0.15\mu$m, $N = 1980$. Total area: $A = 112.86\mu$m^2

V. Results and validation

A. HCI lifetime extrapolation

Aggressive downscaling concerning channel length, junction depth and oxide thickness leads to increase the lateral electric field in the transistor ($E \propto 1/L_{eff}$). In such devices, carriers get a high kinetic energy, reason for which they are called "hot carriers". Due to this high energy, hot carriers can either pass the dielectric energy barrier or generate an electron/hole pair by impact ionization. In both cases a charge may be injected into the dielectric (Hot Carrier Injection), degrading the device reliability (lifetime reduction) and shifting the threshold voltage. HCI lifetime strongly depends on L_{eff}. The latter has to be measured with accuracy in order to distinguish itself from other factors which may affect the lifetime too.

We performed L_{eff} measurements on two devices from 65nm-LPST technology ($T_{ox} \cong 18.5$Å), allowing the use of matrix test structures described in IV. HCI lifetime is expected to be the same because devices come from similar process flows ('A' and 'B'), in which L_{eff} is the only HCI-relevant factor which could change. Indeed reliability measurements show a lifetime-shift which can be explained by a L_{eff}-shift of 4nm between 'A'- and 'B'-processed devices.
L_{eff} measurements have been done using methods described into III-B on more than 20 dies. Accurate and consistent results were obtained: $\Delta L_{eff} = L_{eff}^{A} - L_{eff}^{B} \cong (3.5 \pm 1)$nm, validating the assumption that HCI lifetime shift is exclusively due to a L_{eff} shift between the two devices.

This example clearly shows the relevance of a L_{eff} measurements towards reliability studies and usefulness of L_{eff} extraction for physical understanding of the transistor.

B. Mobility measurements

We performed $I_D(V_G)$ measurements further to L_{eff} extraction in order to extract the mobility and study its behavior toward L_{eff}. We focused on low field mobility ($\mu_0 \approx \mu_{eff}\,(Q_{inv} \sim 0)$) instead of the whole curve $\mu_{eff}(E_{eff})$ obtained by Split C-V method [2]. Thus we need accurate

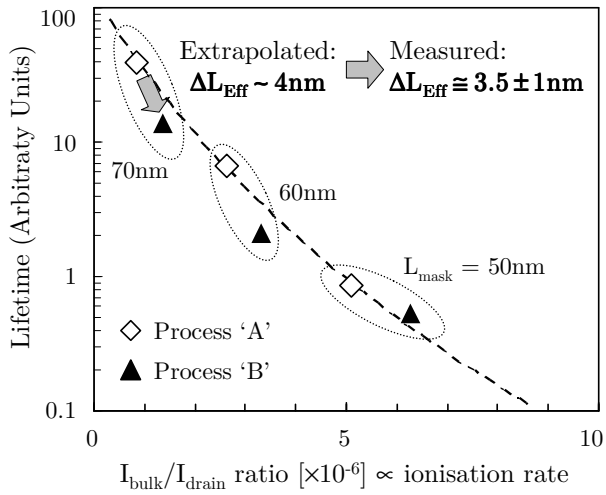

Fig. 7. Lifetime versus ratio I_{bulk}/I_{drain} plot for two different gate annealings. A 4nm-shift is extrapolated from the measurements, for $L_{mask} =$ 70nm.

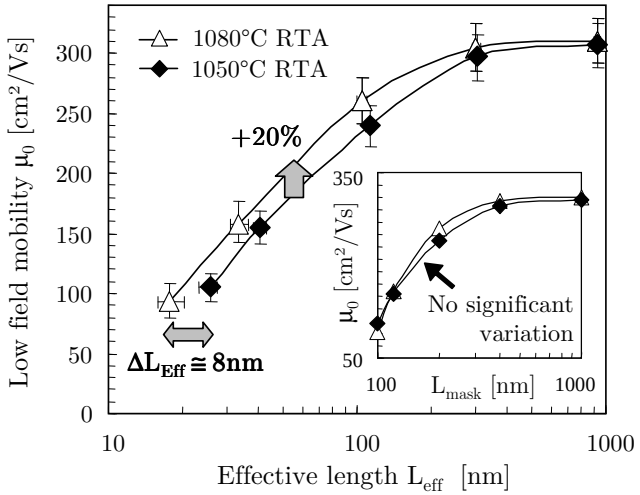

Fig. 8. Mobility behavior toward L_{eff} and L_{mask} (insert) for 45nm-devices ($T_{ox} \cong$ 12Å). $\mu_0(L_{eff})$ plot discloses a 20% mobility improvement on short devices.

L_{eff} extractions in order to extract μ_0 with accuracy too. Isolated transistors are used for $I_D(V_G)$ measurements, instead of matrix structures which exceed the current-compliance of the measurement setup. μ_0 is extracted by coupling β parameter extraction using the Y-function method [12] with L_{eff} extraction described in this paper. Finally, μ_0 is deduced from (4) where β is defined by current equation in linear regime (3).

$$I_D = \beta \cdot V_{DS} \frac{V_G - V_{th} - 0.5V_{DS}}{1 + \theta_1(V_G - V_{th}) + \theta_2(V_G - V_{th})^2} \quad (3)$$

$$\beta = \mu_0 C_{Ox} \frac{W}{L_{eff}} \Rightarrow \mu_0\,(L_{eff}) = \frac{\beta L_{eff}}{W C_{Ox}} \quad (4)$$

We performed $\mu_0(L_{eff})$ extractions on advanced technology (45nm-like, $T_{ox} \cong$ 12Å) processed with two different Rapid Thermal Annealing (RTA) temperatures (1050°C and 1080°C). Fig. 8 compare $\mu_0(L_{eff})$ and $\mu_0(L_{mask})$ plots for both devices to highlight the usefulness of our L_{eff} extraction method in this kind of study. Actually, $\mu_0(L_{eff})$ results show an 8nm L_{eff}-shift in addition to a 20% mobility improvement on short devices between both RTA temperatures. This would have been imperceptible on a $\mu_0(L_{mask})$ plot even by knowing the right μ_0 values (insert Fig. 8).
Mobility-shift induced by annealing temperature disclosed a new physical mechanism which confirmed existence of neutral defects in the channel, near the junctions [7].

VI. Conclusion

We demonstrate high capabilities of our newly industrially-adapted L_{eff} extraction. Results with outstanding accuracy were obtained ($\pm$ 1nm) and offer unequaled benefits towards mobility extraction and HCI lifetime predictions. Systematic and statistical measurements as been done thanks to new matrix test structures, reducing measurement time. This method could be extended to in line monitoring in a near future, to facilitate development of new generation devices.

Acknowledgment

The authors would like to thank the Alliance Crolles2 Advanced Modules and Process Integration teams for providing devices used in this work.

References

[1] A. Scholten, R. Duffy, R. van Langevelde, and D. Klaassen, "Compact modelling of pocket-implanted MOSFETs," in *Proc. of 31st European Solid-State Device Research Conference (ESSDERC'01)*, Nuremberg, Germany, Sep. 2001, pp. 311–314.

[2] K. Romanjek, F. Andrieu, T. Ernst, and G. Ghibaudo, "Characterization of the effective mobility by split C(V) techniquein sub 0.1 lm si and SiGe PMOSFETs," *Solid State Electronics*, vol. 49, pp. 721–726, 2005.

[3] G. Hu, C. Chang, and Y.-T. Chia, "Gate-voltage-dependent effective channel length and series resistance of LDD MOSFET's," *IEEE Trans. Electron Devices*, vol. 34, pp. 2469–2475, Dec. 1987.

[4] Y. Taur, D. Zicherman, D. Lombardi, P. Restle, C. Hsu, H. Nanafi, M. Wordeman, B. Davari, and G. Shahidi, "A new 'shift and ratio' method for MOSFET channel-length extraction," *IEEE Electron Device Lett.*, vol. 13, pp. 267–269, May 1992.

[5] B. Cretu, T. Boutchacha, G. Ghibaudo, and F. Balestra, "New ratio method for effective channel length and threshold voltage extraction in MOS transistors," *IEEE Electron Device Lett.*, vol. 37, pp. 717–719, May 2001.

[6] Y. Taur, "Mosfet channel length: Extraction and interpretation," *IEEE Electron Device Lett.*, vol. 47, pp. 160–170, Jan. 2000.

[7] A. Cros, K. Romanjek, D. Fleury, S. Harrison, R. Cerutti, P. Coronel, B. Dumont, A. Pouydebasque, R. Wacquez, B. Duriez, R. Gwoziecki, F. Boeuf, H. Brut, G. Ghibaudo, and T. Skotnicki, "Unexpected mobility degradation for very short devices : A new challenge for cmos scaling," in *Proc. IEEE Int. Electron Devices Meeting (IEDM'06)*, San Francisco, USA, Dec. 2006, pp. 663–666.

[8] B. Sheu and P. K. Ko, "A capacitance method to determine channel lengths for conventional and LDD MOSFET's," *IEEE Electron Device Lett.*, vol. 5, pp. 491–493, Nov. 1984.

[9] T. Heish, Y.W.Chang, W. Tsai, and T. Lu, "A new leff extraction approach for devices with pocket implants," in *Proc. IEEE Int. Conference on Microelectronic Test Structures (ICMTS'01)*, Kobe, Japan, Mar. 2001, pp. 15–18.

[10] F. Prgaldiny, C. Lallement, and D. Mathiot, "A simple efficient model of parasitic capacitances of deep-submicron ldd mosfets," *Solid State Electronics*, vol. 46, pp. 2191–2198, Jun. 2002.

[11] J. Schmitz, F. N. Cubaynes, R. J. Havens, R. de Kort, A. J. Scholten, and L. F. Tiemeijer, "RF capacitancevoltage characterization of MOSFETs with high leakage dielectrics," *IEEE Electron Device Lett.*, vol. 24, pp. 37–39, Jan. 2003.

[12] G. Ghibaudo, "New method for the extraction of MOSFET parameters," vol. 24, pp. 543–545, Apr. 1988.